Computing
https://doi.org/10.1007/s00607-019-00730-7

Ontology-based discovery of time-series data sources for landslide early warning system

Jedsada Phengsuwan[1] · Tejal Shah[1] · Philip James[2] · Dhavalkumar Thakker[3] · Stuart Barr[2] · Rajiv Ranjan[1]

Received: 1 December 2018 / Accepted: 19 May 2019

Abstract

Modern early warning system (EWS) requires sophisticated knowledge of the natural hazards, the urban context and underlying risk factors to enable dynamic and timely decision making (e.g., hazard detection, hazard preparedness). Landslides are a common form of natural hazard with a global impact and closely linked to a variety of other hazards. EWS for landslides prediction and detection relies on scientific methods and models which requires input from the time series data, such as the earth observation (EO) and urban environment data. Such data sets are produced by a variety of remote sensing satellites and Internet of things sensors which are deployed in the landslide prone areas. To this end, the automatic discovery of potential time series data sources has become a challenge due to the complexity and high variety of data sources. To solve this hard research problem, in this paper, we propose a novel ontology, namely Landslip Ontology, to provide the knowledge base that establishes relationship between landslide hazard and EO and urban data sources. The purpose of Landslip Ontology is to facilitate time series data source discovery for the verification and prediction of landslide hazards. The ontology is evaluated based on scenarios and competency questions to verify the coverage and consistency. Moreover, the ontology can also be used to realize the implementation of data sources discovery system which is an essential component in EWS that needs to manage (store, search, process) rich information from heterogeneous data sources.

Keywords Time series data · IoT data · Early warning system · Landslide hazard · Smart city · High variety data · Ontology · Data sources discovery

Mathematics Subject Classification 68T30

✉ Jedsada Phengsuwan
J.Phengsuwan2@newcastle.ac.uk

Extended author information available on the last page of the article

Published online: 13 June 2019

1 Introduction

The analysis of big time series data has been a grand challenge in several domains including health healthcare [3, 4, 13, 28] and natural hazard management [30]. The advancement of Early Warning Systems (EWS) for natural hazards and urban vulnerabilities is playing a significant role in mitigation and minimising loss of life and damage to infrastructure. EWS systems require strong technical underpinning and sophisticated knowledge of the natural hazards such as the urban context and risk factors to enable dynamic and timely decision-making. Landslides, the main focus of this paper, are a common form of natural hazard that has global importance. Landslides are closely linked with a variety of other natural hazards such as storms, earthquakes, flooding and volcanic eruptions. The prediction of individual landslide occurrence is complex as it depends on many local factors, variables and anthropogenic (caused or produced by human beings) activities. Current EWS for landslides rely on scientific methods such as hyperlocal rainfall monitoring, slope stability models and analysis of remotely sensed images. With the emergence of Internet of things (IoT), decision makers are also analysing observation and measurement data produced by sensors (e.g., soil moisture, soil movement, rainfall, humidity, wind speed) which are deployed in landslide prone areas.

Moreover, the emergence of social media (e.g. Facebook, Twitter and Instagram) has lead to the possibility for general public to also contribute to landslide monitoring by reporting warning signs related to landslide events. Before EWS can optimally utilize information from multiple, heterogeneous time series data sources (e.g., social media, IoT sensors), it is essential to realise a common knowledge base for capturing the core conceptual information and the cross co-relationship between events (that could be potentially discovered by analysing those data sources). Moreover, cross co-analysis of time series data sources is not only useful for the discovery of event co-relation but also allows for additional event verification. For example, landslide early warning sign detected by processing Twitter streams (e.g., by monitoring tweets relevant to landslides) can be verified by analyzing IoT sensor data or other corroborating data (e.g., news feed, remotely sensed satellite data) obtained from the area of interest. However, discovering such cross co-relationship of events from heterogeneous time series data sources has many challenges including lack of common terminology and presence of implicit relationships that are difficult to manually identify and analyse.

The main *contribution* of this paper is a formal knowledge base of landslide domain concepts to enable the integration of time series data from multiple heterogeneous sources for real-time analysis and early prediction of landslide events. Underpinning this knowledge base is the Landslip Ontology that captures the relationships between landslides, multi-hazards, warning signs, sensor data and other time series data sources. The purpose of the ontology is to facilitate data discovery which will be used to find potential data sources for landslide verification. The proposed Landslip Ontology is evaluated based on scenarios and evidence from landslide hazard in Southern India (an area prone to landslide activity) [16]. The experimental results show the accuracy of the data discovery mechanism and indicate the benefits of using social media (along with other time series data sources) as a potential warning mechanism to bolster the potential of landslide early warning.

The rest of this paper is organized as follow: related work is discussed nest, followed by a Landslip scenario in Sect. 3. The detail of Landslip Ontology is described in Sect. 4, followed by the design of data sources discovery system in Sect. 5. The evaluation of Landslip Ontology is discussed in Sect. 6. Finally, we conclude our paper and future work in Sect. 7.

2 Related work

2.1 Data utilisation in multi-hazard early warning system

Multi-hazard refers to a collection of multiple major hazards that a country faces [30]. There is a possibility that several hazardous events occur simultaneously and are interrelated. Tropical storms, for example, is one of the most common environmental hazards (in the tropics), which can trigger multiple hazards such as heavy rainfall that in turn can induce flash flooding. Furthermore, heavy rain and flooding increase the moisture content of soil in a mountainous area and this may induce landslide. To minimise the loss of life and property damage from these inter-related hazards a comprehensive strategy for hazard management is required. In general, a strategy for hazard management is comprised of four phases [32]: (i) *mitigation* the actions to minimise the cause and impact of hazards and prevent them from developing into full-blown disaster; (ii) *preparedness* the action plans and educational activities for communities to confront with unpreventable hazard events; (iii) *response* the actions for emergency situations to protect people life and properties during hazard or disaster events; and (iv) *recovery* the actions to restore damaged properties and communitys infrastructures and to cure people from their illnesses. These four phases demand supporting tools and technologies to enhance the effectiveness of hazard management.

Several modern multi-hazard early warning systems take advantage of the data explosion on the social media. Authors in [29] proposes using a twitter data analysis framework for identifying tweets that are relevant to a particular type of disaster (e.g. earthquake, flood, and wildfire). Several techniques, including matching-based and learning-based, to identify relevant tweets are also evaluated. The work in [14] studies the potential of using social media data to identify peatland fires and haze events in Sumatra Island, Indonesia. A data classification algorithm is used to analyse the tweets and the results are verified by using hotspot and air quality data from NASA satellite imagery. A data classification algorithm is also used in [26] to automatically classify tweets and text messages (from Ushahidi crowdsourcing application) generated during the Haiti earthquake in 2010. The goal of their work is to provide an information infrastructure for timely delivery of appropriately classified messages to the appropriate responsible departments. Work in [12] proposed a decision support system that integrates crowd sourcing information with Wireless Sensor Networks (WSN) to improve the coverage of monitoring area in flood risk management in Brazil. This research introduces the Open Geospatial Consortium (OGC) standards to facilitate the data integration among crowd sourcing information and WSN.

2.2 Semantic web technologies and high variety data management for multi-hazards

Earth Observation (EO) and urban data provided by multiple data sources are accessible by different methods ranging from direct download to various standard Web Services APIs (e.g. Web Map Services, Web Feature Services, Sensor Observation Services, RESTful API, SOAP-based API, etc.). In addition, there are heterogeneities among EO and urban data provided by different data sources [7] including: (i) syntactic heterogeneity the difference in data format or data model for presenting datasets (e.g. plain text, CSV, Excel, XML, JSON, O&M, SensorML, etc.); (ii) structural heterogeneity the difference in data schema for describing the same types of datasets (e.g. describing soil moisture using different XML Schemas); and (iii) semantic heterogeneity difference in meaning or context of the content in datasets. These heterogeneities reveal the challenging problems brought forth by the high variety data availability in multi-hazard applications. Semantic Web Technologies have thus played a significant role by providing languages and tools for modelling domains including describing the concept and relationship among the data and hazardous events. According to W3C definition [35, 36], the Semantic Web is a web of data that provides a common framework for data sharing and reuse across applications, enterprises, and communities.

Ontology, a key element of the Semantic Web, is a specification of a conceptual model for describing knowledge about a domain of interest. A basic concept in a form of ontology can be described by an Resource Description Framework (RDF) triple [33] which is comprised of a subject, a predicate and an object. Concepts described by RDF can be extended by Web Ontology Language (OWL) [34] to construct an ontology for representing rich and complex knowledge about things. In the case of multi-hazards application, an ontology can be used to: (i) represent domain knowledge through concepts, their attributes and relationships between data sources, data and hazards; and (ii) facilitate data integration across multiple data sources that represent varieties, velocity and volume characteristics of Big Data.

Ontologies are widely used in hazard management to model knowledge about hazards and use it to manage actual data derived from EO and urban sources. Hazard assessment and urbanisation analysis are two of the common application areas where ontologies are used. The Semantic Sensor Network Ontology (SSN) [22] and the Semantic Web for Earth and Environmental Terminology (SWEET) [25] are two significant ontologies that are commonly applied for hazard management. Authors in [22] reuse SWEET to conceptualize knowledge and expertise of several areas, such as buried assets (e.g. pipes and cables), soil, roads, the natural environment and human activities. Additionally, the Ontology of Soil Properties and Process (OSP) is proposed in their work to describe a concept of soil properties (e.g. soil strength) and process of soil (e.g. soil compaction). The OSP and other concepts are used to express how they affect each other in asset maintenance activities. Furthermore, [2] and [22] present the application of SSN for wind monitoring. The first work uses SSN with Ontology for Kinds and Units (QU) [18] to conceptualise wind properties (e.g. wind speed and direction) while the later uses SSN and SWEET to model the concepts of wind sensors and data streams of wind observations. The Landslides ontology [5] extends SSN

to organized knowledge for the landslides domain such as the concepts of landslides, earthquake, geographical units, soil, precipitation and wind. Even though these ontologies provide comprehensive concepts for sensor data and hazard event, and provide a reusable, widely used semantic underpinning, they do not cover conceptual aspects on human sensors (e.g. social media data). Hence, currently additional processes are required when applying these ontologies to EWS for multi-hazard application.

The related literature in the context of multi-hazard management can be classified based on the following three perspectives, *data sources*, *hazardous event analytics*, and *EO and urban time series data management*. It can be seen that effective multi-hazard management demands high quality and rich data from vast amount of data sources that are related to the hazard of interest. Data sources utilized by multi-hazard management applications can be any sensors and/or data services that provide EO and urban data. Such data sources include *physical sensor* (e.g. remote sensing, in situ sensor, wireless sensor network) and *human sensor* (e.g. social media, blogs and crowd sourcing). Recent data analytics research for multi-hazard management focused on hazardous event analysis, which are conducted into three main directions, *event identification*, *event verification*, and *event prediction*. These research reveal the challenging problems in the EO and urban time series data management, especially the discovery of potential time series data sources over the complexity and high variety of such data sources in multi-hazard management applications. Ontology is a common method for not only modeling knowledge about hazard but also managing EO and urban data. Recent work around developing the ontology in this domain are classified as *standardizing ontology* and *reusing ontology*. They have shown that current standard ontologies for data sources discovery do not exist. In addition, existing applications of ontology in this domain mostly investigate specific problems, in other words these approaches are not generalized. They fail to model the relationship between data sources and the domain knowledge which is an important factor for efficient data integration and data sources discovery.

3 Landslip scenario

The development of Early Warning System and Decision Support System for multi-hazards can be accomplished in several approaches [17], depending on (i) the rules stakeholders engage in hazard risk reduction, (ii) geographical conditions of hazard prone area, and (iii) EO and urban data provided by responsible organizations. To achieve the goals of risk prevention and mitigation on landslide multi-hazards, scenario-based approach [15] is thus used in order to specify the scope of landslide problems and landslide hazard management activities. Additionally, the scenario-based approach is defined as a narrative story that represent expected uses of a system in the domain of interest from both domain experts and ontology developers viewpoints. Therefore, the scenario helps to identify the scope of the domain ontology to be designed.

3.1 Scenario

The Landslip scenario is co-created with domain experts who are members of Landslip. In addition, the Landslip is an NERC funded project involving the analysis of observation data and social media data provided by several institutions to contribute to the reduction of landslide impacts in the risk area of India. The scenario focuses on the *preparedness phase* of disaster management where *warning signs* of landslide from social media and observation data are detected before the occurrence of landslides. The landslide warning sign is an incident that indicates the potential of landslide hazard. It can be observed by human or an early warning system. The examples of landslide warning signs are the physical changes of utilities or infrastructure (e.g. blocked road, Leaning telephone poles, retaining walls or fences), a movement of soil from foundation, and a change of color in a river. In addition, a warning sign observed by people who are unknowledgeable about landslide hazard (e.g. social media users) requires additional process to verify the potential of landslide. Designing the scenario, historical event of landslide and domain experts experiences are considered. Figure 1 illustrates a situation before landslide that happens in a place located in an urban area. This area encompasses both natural environment (e.g. river, and mountain) and built environment (e.g. schools, hospitals, road, water supply and electricity). Locating in a high slope area, the place is prone to landslide and is monitoring by the National Disaster Management Authority (NDMA). Here, an expert from NDMA analyses satellite images to detect warning signs and informs decision makers for the potential landslide hazard. Meanwhile, *person A* who lives in the place is enjoying his leisure time by walking around his house. Living in the landslide prone area, it raises his awareness on possibility of landslide hazard. Then he is keen to contribute to his community by reporting any incident observed in his daily life via his social media account. While walking around his place, he has observed a leaning pole nearby his house. Thus, he takes a photo of the leaning pole and reports the incident to social network using his Twitter account. He also gives additional information such as observed time and place to the tweet. Besides, *person B* who lives in a nearby area has observed that the color of tab water in his house become brown. He thus uses hist Facebook account to report this incident. These messages from social media are collected by NDMA Early Warning System (EWS) to detect landslide warning signs. Here, an NDMA's staff who is a member of decision making team receives a notification message from EWS with regard to the leaning telephone pole. The incident is considered as a warning sign for landslide hazard. Receiving such information from social media users, the decision maker needs to verify the information before making further decision. Base on this, the decision maker who is a domain expert in landslide hazard risk assessment performs data analysis using an appropriate landslide analytical model. This process requires adequate historical events of landslide, EO and urban data provided by various data sources to get more accuracy of the data analysis. Hence, the decision maker searches for potential data sources from the Data Sources Discovery Service (DS) and gathers EO and urban data from the discovered data sources. The gathered data is then used to verify the social media information. Furthermore, the decision maker utilizes the EWS to assess the risk and impact of the landslide event using EO and urban data

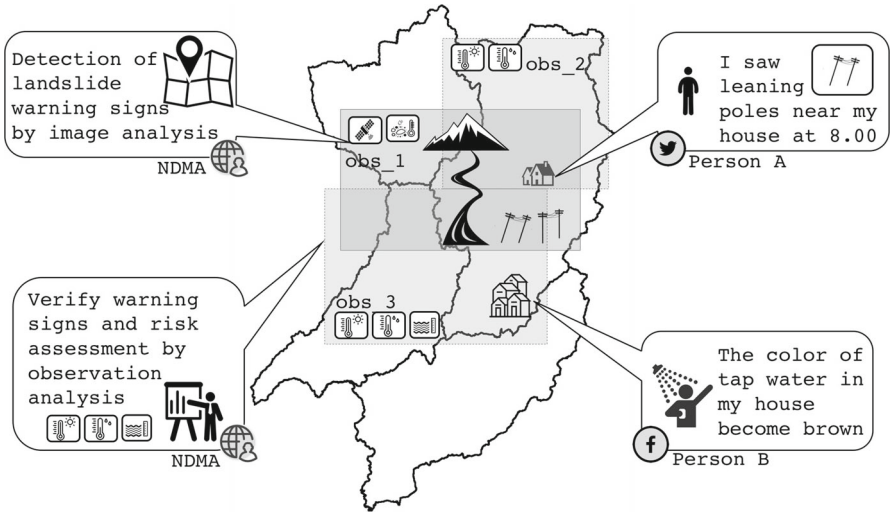

Fig. 1 A Scenario of Landslide Multi-hazard

and take timely actions against the event. An example of such actions is disseminating actionable warning information to people in the landslide prone area.

3.2 Overall concepts

The above scenario reveals the essential role of data-driven early warning system for landslide hazard management which comprises of 5 main components.

- *Exposure*—refers to people and environment which are living or located in landslide hazard prone area and are affected by landslide multi-hazard. In addition, environment can be classified to natural environment and built environment. The natural environment is all living and non-living things that occurred naturally (e.g. animals, river, forest, mountain, etc.). On the other hand, the built environment [6] is a combination of infrastructures and facilities produced by people as a core foundation in the community (e.g. house, school, road, bridge, electricity, water supply, etc.)
- *Stakeholder*—refers to people or organizations who have a stake in the landslide event. In the scenario, stakeholders are: (i) *social media users* who report landslide warning signs through their social media (e.g. Facebook, Twitter, Instragrams, etc.); (ii) *data collectors and providers* who deploy sensor devices in landslide hazard prone area and provide EO and urban data collecting from such devices to EWS for analysis. Data providers also include the third parties who collect data from sensor devices owned by the others. (iii) *Decision makers* who have responsible for conducting landslide hazard risk assessment using available social media data and EO and urban data. They make a decision based on result from Decision Support System and hazard risk management plan in order to inform people in risk area before the occurrence of landslide hazard.

- *Event*—refers to an occurrence which is related to a hazard. Additionally, hazard itself is also consider as an event. The hazard-related event is classified as pre-hazard event, post-hazard event and event during hazard. Since Early Warning System analyse EO and urban data to predict the potential of hazard in the area of interest, *warning signs* and *anthropogenic processes* are the majority of events in this scenario. In addition, a warning signs is an event that can indicate a possibility of hazards. An example of the warning sign is broken underground utilities which can be a warning sign for landslide hazard. An anthropogenic process refers to human activities which can induce hazards. An example of such activities is vegetation removal which induce landslide.
- *Data Sources*—refer to any sensors and data services that provide data to data consumers. These data sources have different capabilities to provide data. Sensor is a component that observes and measures physical phenomena and transform the observation and measurement into a human readable form. There are two types of sensor, *physical sensor* and *human sensor*. The data service is an application software that collects, stores and provides data from multiple devices. Several types of data sources are currently available to provide EO and urban data for multi-hazard applications.
- *Decision Support Applications*—refer to an integrated system that provide functionalities for stakeholders to monitor, forecast and predict, validate and assess hazardous events. In this scenario, EO and urban data collection system, data sources discovery services, hazardous event detection system and Early Warning System (EWS) are major components of Decision Support Applications. As a consequence, these applications enable stakeholders to take timely actions to reduce impacts of landslide hazard in advance. For example, once a landslide hazard is likely to be happened, a decision maker can make a decision based on information and knowledge from EWS to disseminate actionable warnings information to people in the landslide prone area.

The Data-driven early EWS analyses landslide-related data to enable dynamic and timely decision making against landslide hazard. Such data includes historical landslide events, historical and real-time observation data generated by physical sensors, and social media data. In addition, a number of sensor devices have been deployed in the landslide hazard prone area by organizations who are in charge of landslide hazard management. the organizations collect EO and urban data from their sensors to monitor landslide hazard events in real-time. Besides, the collected data is stored in their local repositories and is provided as data sources to their co-ordinated organizations for further analysis. Here, the data sources metadata is published to a *Data Sources Discovery Services* (*DS*) which is an application of Decision Support System. The DS enables data publishers to advertise their data sources by registering data sources metadata to a metadata registry service. Moreover, It allows data consumers to search for their potential data sources to be used in their applications.

4 Landslip Ontology

Our proposed Landslip Ontology is developed based on NeOn [27] methodology. We define scope and purpose of the ontology based on the scenario mentioned in Sect. 3. The ontology is implemented in OWL and is built in Protege. According to the scenario mentioned in the previous section, Landslip Ontology is designed and developed to conceptualize the knowledge of landslide hazard and its warning signs. Moreover, knowledge of data sources is also necessary to facilitate data sources discovery and landslide precursor verification. Based on this, Landslip Ontology is comprised of two main modules, Landslip Common and Landslip-DataSources.

Scope and Purpose The development of Landslip Ontology is driven by the goal of Landslip project to mitigate the impact of landslide hazard. Thus, the ontology focuses on the preparedness phase of disaster management where landslide warning signs play an important role to indicate the potential of landslide. The scope of Landslip Ontology is defined based on the scenario in Sect. 3. Based on this, the ontology conceptualizes knowledges of landslide hazard specifically causes of landslide hazard and multi-hazards interactions which can trigger landslide hazard. Furthermore, the ontology conceptualizes landslide-related incidents which can be observed by people in landslide prone area. These incidents are considered as warning signs for landslide hazard.

The concepts of landslide hazards are linked to EO and urban data which are set of properties for landslide observation. The ontology focuses on Landslide multi-hazard domain. The level of granularity is determined to the competency questions and terms identified.

Knowledge Sources The ontology is design based on knowledge and experiences from four scientists and experts, from Landslip project, who are specialists in landslide hazard management with average 10 years experiences. Specifically, One scientist works for British Geological Survey (BGS) with focus on multi-hazard management. Other One is a scientist from Geological Survey of India (GSI) who are working on landslide hazard management in India. The two others are academic staffs who are specialist in natural hazard and geoscience.

Besides, publications [8, 9] and standard specifications [19–22, 25] involving multi-hazards and geo-spatial data models are also used as additional knowledge sources to design the ontology.

Figure 2 depicts overall concepts of our proposed Landslip Ontology. The ontology is comprised of two modules: (i) *Landslip Common Ontology*—defines concepts about landslide hazard and its interaction to another hazards and anthropogenic process; and (ii) *Landslip Data Sources Ontology*—defines concepts about observation and data sources for landslide hazard risk assessment. The Landslip Ontology reuses SSN ontology and terminology defined in OGC standards (e.g. Observation and Measurement [19], SensorML [20] and SOS [21]).

4.1 Landslip Common Ontology

The purpose of Landslip Common Ontology is to provide a conceptual knowledge model of landslide domain. In addition, the Common Ontology is a combination of the-

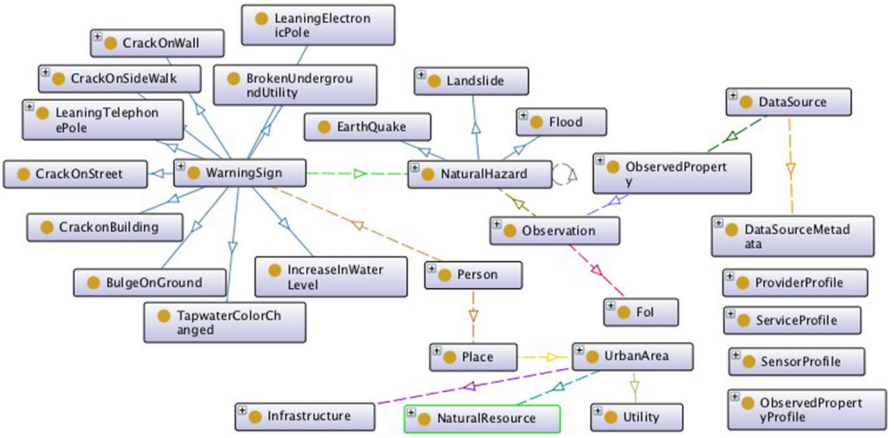

Fig. 2 Overall of Landslip Ontology

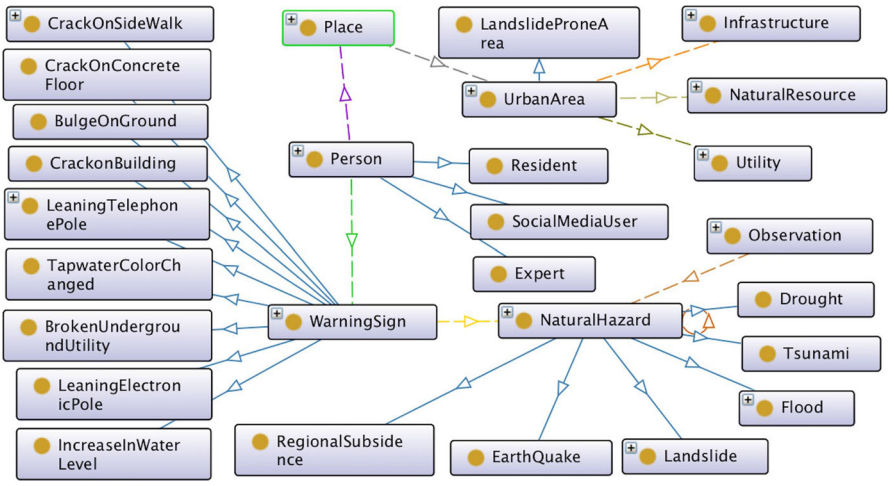

Fig. 3 Common Ontology

oretical knowledge and human experiences to identify warning sign before landslide. Basically, landslide is one of the most significant multihazards which can be found in many places around the globe [11]. Such hazard has interactions or can be triggered by another hazards [8]. Based on this, the Landslip Common Ontology conceptualizes knowledges of landslide and its interaction with other multi-hazards [8, 9] and knowledge of warning signs that can be observed by human and use such knowledge to indicate landslide event before the occurrence of landslide. The knowledge defined in the ontology can be used to facilitate landslide early warning based on warning signs observed and reported by people in social network. Figure 3a illustrates the concept of the Landslip Common Ontology which comprises of four main concepts as follow:

– *UrbanArea*—defines concepts about urban area that prone to landslide including its basic elements. The urban area encompasses both natural resources (e.g. river, and

mountain) and built environment, including infrastructure (e.g. road and railway), utility (e.g. electricity and tab water) and place (e.g. school, hospital, house and flat). Located in landslide prone area, these elements can be affected by landslide and other multi-hazards.
- *NaturalHazard*—defines a set of multi-hazards which can trigger land-slide hazard. This concept captures knowledge mainly on the interactions between landslide hazard and other multi-hazards (e.g. flood, earthquake, tsunami, and drought). In addition, the interactions between other multi-hazards are able to indicate landslide hazard.
- *AnthropogenicProcess*—defines a set of human activities that produce negative effects to landslide [8]. This concept also captures knowledge about the interaction with in the processes to provide direct and indirect indications of landslide hazards. In addition, the direct indications are the processes that are a trigger of landslide while the indirect indications are the processes that trigger other processes which trigger landslide. Moreover, the major indicators for anthropogenic processes are warning sign observed by a person.
- *WarningSign*—defines a set of incidents that can be an indications of landslide hazard, other multi-hazards and anthropogenic processes. The concept of warning sign is mainly focus on incidents which can be observed by a person. Such incidents are useful for landslide EWS in order to detect landslide precursors based on incidents reported in social network.

4.2 Landslip data sources ontology

EO and urban data observed by sensors indicate events or changes of landslide phenomena [24]. Such data (e.g. rain, temperature, soil moisture. etc.) from a variety of data sources is collected by data provider and provides for stakeholders to be used in their landslide hazard applications [23]. Due to the high variety and geographically distributed nature of OE and urban data sources, effective data sources discovery [37] is thus required in order to provide sufficient amount and quality of data for landslide hazard risk assessment. Landslip Data Sources Ontology is developed to enable semantically discovery of data sources. In addition, the ontology describes concepts and relationships of EO and urban data, data sources, sensor devices, and data providers. With the combination of this ontology and Landslip Common Ontology, data sources discovery mechanism utilizes knowledge of landslide hazards to discover data sources which are related to the hazard of interest. Specifically, the knowledge of landslide warning sign can be used to identify appropriate observed properties and data sources for landslide precursor verifications. This capability enable EWS to provide dynamic and timely decision making against landslide hazards. Figure 4b illustrates the Landslip Data Sources Ontology which comprises of three main concepts and reuse SSN Ontology [22] and OGC standard [19–21] for the concepts of observation and sensors.

- *DataSource*—is the main concept of Landslip Data Sources Ontology. A data source is any sensors or data services that provide observation data (e.g. physical sensor, human sensor and data service). DataSource defines a set of comprehensive infor-

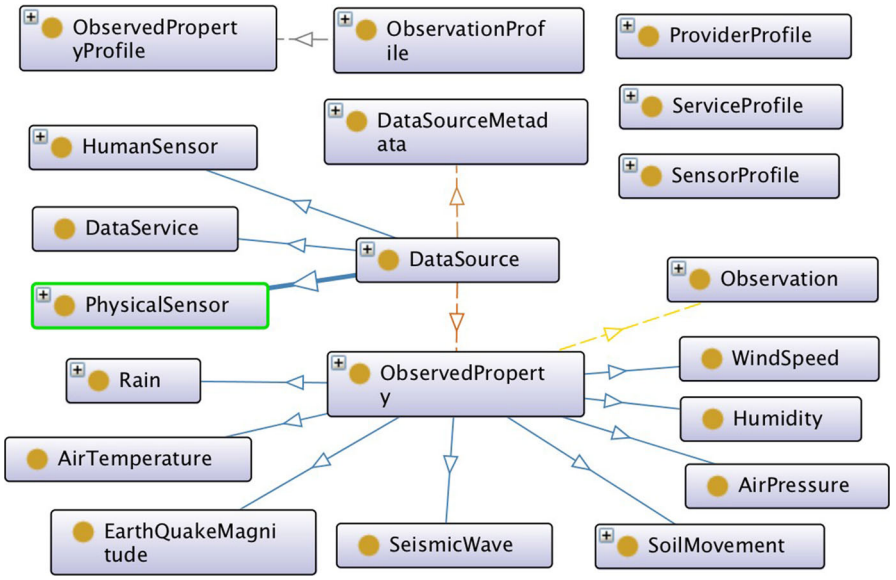

Fig. 4 Data Source Ontology

mation related to observation and data sources metadata which are the details of data sources.

– *Observation*—defines a set of observed properties (EO and urban data) which are used to observed features of interest related to landslide hazard. The examples of observed properties are soil moisture, soil movement, rain, earth quake magnitude, temperature, humidity, and wind speed. These observed properties are accessible to EWS via data sources.
– *DataSourceMetadata*—defines a set of information which are necessary for data acquisition process. This concepts is comprised of four groups of information profile: (i) observation profile—a set of observed properties provided by a data source; (ii) observed property profile—provides information about data type, feature of interest, and phenomenon time; (iii) sensor profile—provides information about type of sensor, feature of interest, and list of event to be observed; (iv) service profile—provides information which can be used to access a service (e.g. service type, end-point, provider); and (v) provider profile—provide the information about data provider (e.g. provider name, contact address).

4.3 Ontology metrics

Table 1 shows a summary of the ontological features of Landslip Ontology in terms of size (number of classes, properties, and individuals), expressivity, and complexity of the core knowledge captured by axioms.

Table 1 Landslip Ontology features

Feature	Value
No of classes	98
No of properties	26
No of individuals	30
No of axioms	462
DL expressivity	ALCH(D)

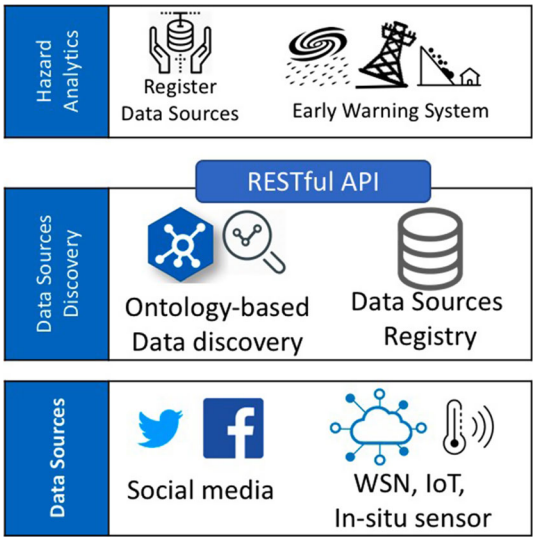

Fig. 5 Landslip Data Sources Discovery Service architecture

5 System architecture

To realize the ontology-based data sources discovery system, we have designed the architecture which comprise of three main layers: (i) data sources layer; (ii) data discovery layer; and; and (iii) hazard applications layer. Figure 5, depicts the overview architecture of our proposed data sources discovery services.

- *data sources layer*—consist of a number of data sources provided by various data providers. Data sources collects EO and urban data from physical sensors deployed in landslide prone area. These sensors observe or measure properties of landslide and other earth observation which can be use to indicate landslide hazard. The data sources are accessible through a variety of methods (e.g. REST API, RDBMS, WSN) depending on data source providers. Moreover, data from social medias is also considered as data sources in this layer.
- *data sources discovery layer*—maintains the Landslip ontology which represents knowledge of landslide and data sources in a triplestore. It also provides data sources registry which store data sources metadata, including metadata for sensor, service and observation. Furthermore, there are a number of functionalities provided by this layer which allow users to (1) publish data sources; (2) search for potential data

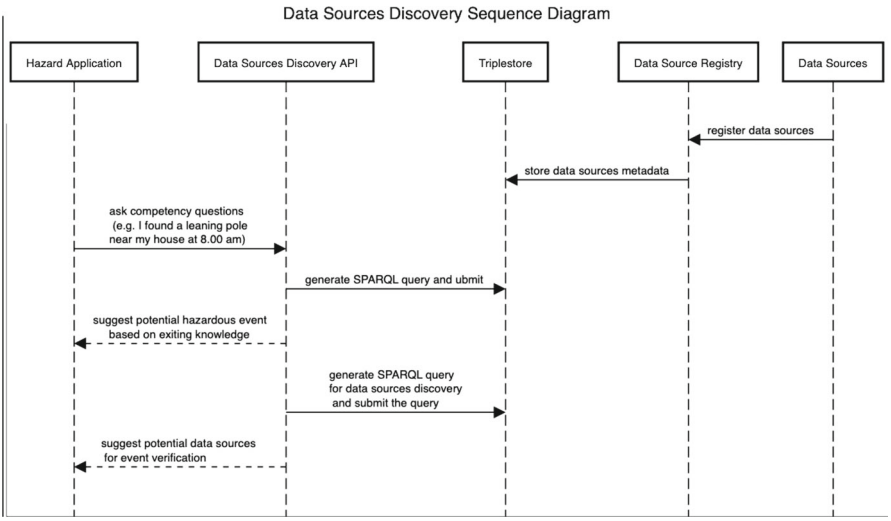

Fig. 6 Data Sources Discovery Sequence Diagram

sources; and (3) indicate landslide hazard using warning signs. These functionalities are accomplished based on knowledge of landslide and data sources provided by the Landslip Ontology. In addition, the functionalities provided by this layer is accessible through data sources discovery service APIs which are available in form of RESTful Web Services API.

– *hazard applications layer*—provides client APIs to access the functionalities offered by the data sources discovery layers. In addition, the client APIs are design for both data provider and data consumer. Here, data provider can user the client API to register their data sources along with data sources metadata. On the other hand, data consumer uses the client API to search for potential data sources based on landslide warning sign.

Figure 6 illustrates the interactions among the three layers. Initially, multiple data sources provided by different providers are registered to the data sources registry. In addition, the actual knowledge of landslide is constructed based on Landslip Ontology and information extracted from social media. Both data sources metadata and landslide knowledge are stored in Triplestore. Here, a hazard application utilizes the system by invoking the Data Sources Discovery API to ask a competency question which is related to landslide multi-hazard. The API then generates a SPARQL query which correspond to the selected competency question and submit the Triplestore for reasoning query. As a result, the API suggests potential hazardous event based on existing knowledge. Built from social media, the knowledge requires further analysis to verify the correctness of the suggested event. The API generates additional SPARQL query for the discovery of potential data sources. Finally, data sources metadata providing the detail of the potential data sources is returned to the hazard application. The application then use the information to access actual data sources and retrieve EO and urban time series data for hazard event verification and other data analytics.

Table 2 An example of competency questions

	Competency questions
Q1	What other hazards are likely to happen when hazard H has happened?
Q2	What is the probability of an event E occurring when warning sign W has been observed?
Q3	What is the probability of an event E occurring when a set of warning sign, W_1, W_2, W_3, W_n have been observed?
Q4	Is warning sign W an indicator for landslide L?
Q5	What are observed properties that can be used to verify landslide when a warning sign W is observed?
Q6	Identify the data sources and their metadata required to observe a set of hazards (H_1, H_2, H_3, H_n)

6 Evaluation

An evaluation was conducted to verify the coverage of the Landslip Ontology and its application in landslide early warning. Whilst various approaches for evaluating an ontology exist, competency questions remain the most common approach [1, 31]. This approach stipulates that an ontology must be able to represent the competency questions using its terminology and answer these questions using the axioms [10]. According to the use case mentioned Sect. 3, some of competency questions are developed as shown in Table 2. We arranged an interview with domain experts who are members of the Landslip project. Those domain experts include two academic staffs who are specialist in natural hazard and geoscience, and a scientist from British Geological Survey (BGS). Based on the interview, 12 competency questions were received and some of main competency questions are developed as shown in Table 2 for the evaluation.

The competency questions were used for ontology validation and evaluation. The evaluation was conducted using a set of synthesized data that represent the use case of landslide hazard mentioned Sect. 3. We manually added information of natural hazards and EO and urban data to our knowledge base. The information includes landslide hazard, hazard triggers, warning signs, EO and urban data, and data sources. We performed validation over the dataset using Pellet to check for ontology consistency, concept satisfiability, classification, and realisation. Based on the competency questions, we performed preliminary experiments by querying over the knowledge base.

In order to write competency questions and to demonstrate that the LAND-SLIP ontology can be used to ask and answer these questions we use the semantic query language called SPARQL Protocol and RDF Query Language (SPARQL).Using SPARQL query language, we defined query for each competency question to get answers from our knowledge base. Figures 7 and 8 show snapshots of SPARQL query for *Q2* and *Q6* and output for the competency question *Q6* on running the query in protg. By executing the query based on the competency questions *Q1–Q6*, we could verify the coverage of the Landslip Ontology. From the results it can be seen that the ontology is able to identify hazard events based on an observed warning sign. Furthermore, the

Q1

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX :
<http://www.semanticweb.org/ncl/ontologies/2018/6/landslip#>

SELECT ?hazard
         WHERE { :landslide_2 :triggers ?hazard }
```

Q2

```
SELECT ?hazard
WHERE { :leaning_telephone_pole_1 :isWarningSignFor ?hazard }
```

Q3

```
SELECT ?warningSign ?hazard
WHERE { ?warningSign :isWarningSignFor ?hazard .
            VALUES (?warningSign) {
               (:crack_on_wall_2)
               (:crack_on_sideWalk_2)
            } .
}
```

Q4

```
ASK WHERE { :leaning_telephone_pole_1 :isWarningSignFor
:landslide_1 }
```

Q5

```
SELECT    ?observedProperty
WHERE {
         ?observation :isObservationFor :landslide_1 .
         ?observedProperty :isObservedPropertyFor ?observation .
}
```

Q6

```
SELECT  ?hazard ?observation ?observedProperty ?dataSource
?metadata ?profile ?p ?value
WHERE {
         ?observation :isObservationFor ?hazard .
         ?observedProperty :isObservedPropertyFor ?observation .
         ?dataSource :isDataSourceFor ?observedProperty .
         ?dataSource :hasDataSourceMetadata ?metadata .
         ?metadata :hasProfile ?profile .
         ?profile ?p ?value .
         VALUES (?hazard) { (:flood_1) (:landslide_1) } .
         FILTER (?p != rdf:type)
}
```

Fig. 7 SPARQL query for competency question Q1–Q6

ontology can suggest potential data sources and their metadata which can be used by domain experts to perform timely decision making against hazards.

7 Conclusions and future work

Effective Early Warning System (EWS) for Landslide hazard relies on a comprehensive set of EO and urban data provided by geographically distributed data sources. In this paper, we have demonstrated the application of ontology in landslide domain. We proposed a Landslip Ontology to provide the knowledge base and relationship between landslide hazard and data sources. The Landslip Ontology enhance the effective landslide EWS by providing a knowledge base to detect landslide precursor based on incidents reported in social media. Moreover, the ontology enable decision maker

hazard	observation	observedProperty	dataSource	metadata	profile	p	value
flood_1	obs_2	rain_1	dataSource_3	ds_metadata_3	sensor_profile_2	featureOfInterest	"foi_karela_bbox_1"@
flood_1	obs_2	rain_1	dataSource_3	ds_metadata_3	sensor_profile_2	sensorType	"in-situ"@
flood_1	obs_2	rain_1	dataSource_3	ds_metadata_3	sensor_profile_2	eventList	"Flood"@
flood_1	obs_2	rain_1	dataSource_3	ds_metadata_3	provider_profile_2	providerName	"MetOffice"@
flood_1	obs_2	rain_1	dataSource_3	ds_metadata_3	obs_profile_21	observedPropertyType	"Rain"@
flood_1	obs_2	rain_1	dataSource_3	ds_metadata_3	obs_profile_21	featureOfInterest	"foi_karela_bbox_2"@
flood_1	obs_2	rain_1	dataSource_3	ds_metadata_3	obs_profile_21	phenomenonEndTime	"2018-07-21T00:00:00"
flood_1	obs_2	rain_1	dataSource_3	ds_metadata_3	obs_profile_21	phenomenonBeginTime	"2004-01-01T00:00:00"
flood_1	obs_2	rain_1	dataSource_3	ds_metadata_3	obs_profile_21	observedPropertyName	"rain_1"@
landslide_1	obs_1	soil_moisture_1	dataSource_1	ds_metadata_1	service_profile_11	serviceURL	"http://127.0.0.1/rest/m
landslide_1	obs_1	soil_moisture_1	dataSource_1	ds_metadata_1	service_profile_11	serviceProvider	"Amrita"@
landslide_1	obs_1	soil_moisture_1	dataSource_1	ds_metadata_1	service_profile_11	serviceAdapter	"Rest_adaptor_11"@
landslide_1	obs_1	soil_moisture_1	dataSource_1	ds_metadata_1	service_profile_11	serviceType	"REST"@
landslide_1	obs_1	soil_moisture_1	dataSource_1	ds_metadata_1	sensor_profile_1	featureList	"foi_karela_bbox_1"@
landslide_1	obs_1	soil_moisture_1	dataSource_1	ds_metadata_1	sensor_profile_1	eventList	"Landslide"@
landslide_1	obs_1	soil_moisture_1	dataSource_1	ds_metadata_1	sensor_profile_1	sensorType	"in-situ"@

Fig. 8 SPARQL output for competency question Q6

to find potential data sources to verify such incidents. We have performed the evaluation by verifying the coverage of the ontology based on some competency questions. The preliminary experiment over a set of synthesized data have shown the consistency, concept satisfiability, classification and realisation of the ontology. We have also designed the architecture of ontology-based data sources discovery system to realize our proposed ontology. This system became an essential component that allows EWS to discover sufficient data sources and access rich information from the data sources. Our immediate future work on this research is to evaluate the Landslip Ontology in the real application of landslide EWS in Southern India under the Landslip project.

Acknowledgements This research is partially supported by two Natural Environment Research Council projects including LandSlip (NE/P000681/1) and FloodPrep (NE/P017134/1).

References

1. Almeida MB, Barbosa RR (2009) Ontologies in knowledge management support: a case study. J Am Soc Inf Sci Technol 60(10):2032–2047. https://doi.org/10.1002/asi.21120
2. Calbimonte JP, Jeung H, Corcho O, Aberer K (2011) Semantic sensor data search in a large-scale federated sensor network. In: Proceedings of the 4th international workshop on semantic sensor networks, vol 839, pp 23–38
3. Chen D, Hu Y, Wang L, Zomaya AY, Li X (2017) H-parafac: hierarchical parallel factor analysis of multidimensional big data. IEEE Trans Parallel Distrib Syst 28(4):1091–1104. https://doi.org/10.1109/TPDS.2016.2613054
4. Chen D, Li X, Wang L, Khan SU, Wang J, Zeng K, Cai C (2015) Fast and scalable multi-way analysis of massive neural data. IEEE Trans Comput 64(3):707–719. https://doi.org/10.1109/TC.2013.2295806
5. Envision (2016) Ls ontology. http://envision.brgm-rec.fr/LS-Ontologies.aspx. Accessed 15 Nov 2018
6. Gaillard JC (2011) Hazards and the built environment: attaining built-in resilience. Disaster Prev Manag Int J 20(2):215–216. https://doi.org/10.1108/09653561111126148
7. George D (2006) Understanding structural and semantic heterogeneity in the context of database schema integration. In: Proceedings of 6th conference in the department of computing (Journal of the Department of Computing, UCLan), vol 4, pp 29–44
8. Gill J, Malamud B (2016) Hazard interactions and interaction networks (cascades) within multi-hazard methodologies. Earth Syst Dyn 7(3):659–679. https://doi.org/10.5194/esd-7-659-2016

9. Gill JC, Malamud BD (2017) Anthropogenic processes, natural hazards, and interactions in a multi-hazard framework. Earth Sci Rev 166:246–269. https://doi.org/10.1016/j.earscirev.2017.01.002
10. Gruninger M, Fox MS (1994) The role of competency questions in enterprise engineering. In: Proceedings of the IFIP WG5.7 workshop on benchmarking—theory and practice
11. Hong Y, Adler RF (2007) Towards an earlywarning system for global land-slides triggered by rainfall and earthquake. Int J Remote Sens 28(16):3713–3719. https://doi.org/10.1080/01431160701311242
12. Horita FE, Albuquerque JAPD, Degrossi LC, Mendiondo EM, Ueyama J (2015) Development of a spatial decision support system for flood risk management in Brazil that combines volunteered geographic information with wireless sensor networks. Comput Geosci 80(C):84–94. https://doi.org/10.1016/j.cageo.2015.04.001
13. Ke H, Chen D, Shah T, Liu X, Zhang X, Zhang L, Li X (2018) Cloud-aided online eeg classification system for brain healthcare: a case study of depression evaluation with a lightweight CNN. Softw Pract Exp. https://doi.org/10.1002/spe.2668
14. Kibanov M, Stumme G, Amin I, Lee JG (2017) Mining social media to inform peat-land fire and haze disaster management. CoRR http://arxiv.org/abs/1706.05406
15. Knublauch H (2004) Ontology-driven software development in the context of the semantic web: an example scenario with. In: Annex XVII (7)
16. Kuriakose SL, Sankar G, Muraleedharan C (2009) History of landslide susceptibility and a chorology of landslide-prone areas in the western ghats of kerala, india. Environ Geol 57(7):1553–1568. https://doi.org/10.1007/s00254-008-1431-9
17. Lacasse S, Nadim F, Lacasse S, Nadim F (2009) Landslide risk assessment and mitigation strategy, pp 31–61. Springer, Berlin. https://doi.org/10.1007/978-3-540-69970-53
18. Laurent Lefort (CSIRO, A., Group), W.S.S.N.I. (2010) Ontology for quantity kinds and units: units and quantities definitions. https://www.w3.org/2005/Incubator/ssn/ssnx/qu/qu-rec20.html. Accessed 15 Nov 2018
19. OGC (2011) Observations and measurements. http://www.opengeospatial.org/standards/om. Accessed 15 Nov 2018
20. OGC (2011) Sensor model language (sensorml). http://www.opengeospatial.org/standards/sensorml. Accessed 15 Nov 2018
21. OGC (2011) Sensor observation service. http://www.opengeospatial.org/standards/sos. Accessed 15 Nov 2018
22. OGC W (2016) Semantic sensor network ontology. https://www.w3.org/TR/vocabssn/. Accessed 15 Nov 2018
23. Pennington C, Freeborough K, Dashwood C, Dijkstra T, Lawrie K (2015) The national landslide database of great Britain: acquisition, communication and the role of social media. Geomorphology 249:44–51. https://doi.org/10.1016/j.geomorph.2015.03.013 **Geohazard databases: concepts, development, applications**
24. Ranjan R, Phengsuwan J, James P, Barr S, van Moorsel A (2017) Urban risk analytics in the cloud. IT Prof 19(2):4–9. https://doi.org/10.1109/MITP.2017.20
25. Raskin RG, Pan MJ (2005) Knowledge representation in the semantic web for earth and environmental terminology (sweet). Comput Geosci 31(9):1119–1125. https://doi.org/10.1016/j.cageo.2004.12.004 **Application of XML in the geosciences**
26. Shen H (2015) Discussion and analysis of the crowd sourcing mode of public participation in emergency management. In: 2015 8th International symposium on computational intelligence and design (ISCID), vol 2, pp 610–613. https://doi.org/10.1109/iscid.2015.115
27. Súarez-Figueroa MC, Gómez-Pérez A, Fernández-López M (2012) The NeOn methodology for ontology engineering, pp 9–34. Springer, Berlin. https://doi.org/10.1007/978-3-642-24794-12
28. Tang Y, Chen D, Wang L, Zomaya AY, Chen J, Liu H (2018) Bayesian tensor factorization for multi-way analysis of multi-dimensional EEG. Neurocomputing 318:162–174. https://doi.org/10.1016/j.neucom.2018.08.045
29. To H, Agrawal S, Kim SH, Shahabi C (2017) On identifying disaster-related tweets: matching-based or learning-based? CoRR. http://arxiv.org/abs/1705.02009
30. UNISDR (2017) Terminology. https://www.unisdr.org/we/inform/terminology. Accessed 15 Nov 2018
31. Uschold M, Gruninger M (1996) Ontologies: principles, methods and applications. Knowl Eng Rev 11(2):93136. https://doi.org/10.1017/S0269888900007797
32. van Westen C (2012) Landslide risk assessments for decision-making, pp 67–71. The World Bank

33. W3C (2004) Resource description framework (rdf): concepts and abstract syntax. https://www.w3.org/TR/2004/REC-rdf-concepts-20040210/. Accessed 15 Nov 2018
34. W3C (2012) Web ontology language (owl). https://www.w3.org/OWL/. Accessed 15 Nov 2018
35. W3C (2013) W3C data activity building the web of data. https://www.w3.org/2013/data/. Accessed 15 Nov 2018
36. W3C (2013) W3C semantic web activity. https://www.w3.org/2001/sw/. Accessed 15 Nov 2018
37. Wei Q, Jin Z (2012) Service discovery for Internet of things: a context-awareness perspective. In: Proceedings of the fourth Asia-Pacific symposium on Internetware, Internetware'12, pp 25:1–25:6. ACM, New York. https://doi.org/10.1145/2430475.2430500

Affiliations

Jedsada Phengsuwan[1] · Tejal Shah[1] · Philip James[2] · Dhavalkumar Thakker[3] · Stuart Barr[2] · Rajiv Ranjan[1]

Tejal Shah
Tejal.Shah@newcastle.ac.uk

Philip James
Philip.james@newcastle.ac.uk

Dhavalkumar Thakker
D.Thakker@Bradford.ac.uk

Stuart Barr
stuart.barr@newcastle.ac.uk

Rajiv Ranjan
Raj.Ranjan@newcastle.ac.uk

1 School of Computing, Newcastle University, Newcastle upon Tyne, UK

2 School of Engineering, Newcastle University, Newcastle upon Tyne, UK

3 School of Electrical Engineering and Computer Science, University of Bradford, Bradford, UK